Solutions

Vic Titova

The task :

1. Make a 100 bases random DNA sequence with 60% GC content.
2. Make 3 sets of 100 sequences with 2%, 5%, and 10% random errors:

- Insertion - addition of the extra bases
- Deletion - removal of random bases
- Mismatch - replacing one base with a different one

a) Add more errors near the end

b) Add more errors on repetition of the same base(homopolymers) like AA

c) Record the error positions and types. Plot the distribution of error positions for each type.

3. Plot the sequence length distribution for each error rate.
4. Plot the GC content for each error rate.
5. Plot the homopolymer length ratios by base for each error rate.
6. Plot the homopolymer positions by base and error rate.

Expectations, plots, thoughts

Error Position Distribution per Error Type:

- This plot will show where errors are most likely in the sequence.
- I've biased errors towards the end and increased error probability on homopolymers, I expect to see a higher density of errors towards the end of the sequence. Homopolymers should also show a higher error rate.
- To visualise this, a histogram for error type is used. The x-axis represents sequence positions, and the y-axis represents the number or probability of errors.

Sequence Length Distribution per Error Rate:

- This plot will show the distribution of sequence lengths after introducing errors.
- Since insertions add to the sequence length and deletions reduce it, I expect sequences with a 10% error rate to have a wider distribution of lengths than sequences with a 2% error rate.
- A distribution plot is my choice

Expectations, plots, thoughts

GC Content per Error Rate:

- This will show how the GC content changes with different error rates.
- Given that errors are random, the GC content might slightly deviate from the original 60% as the error rate increases, especially as the introduced errors are not biased specifically towards G/C.
- A line plot with an error rate on the x-axis and GC content on the y-axis would be suitable, in my opinion.

Homopolymer Ratios Relative to the Sequence Length per Nucleotide per Error Rate:

- This will show how the ratio of homopolymers (e.g., the ratio of 'AA' sequences to the total sequence length) changes with different error rates.
- Given the increased error rate on homopolymers,I expect the ratio to decrease as the error rate increases because homopolymers might be disrupted more frequently. Perhaps this will be more true for the CG content, as theoretically, there will be more CG homopolymers because of the content.
- A line plot might be suitable, with different colours for each nucleotide type (A, T, C, G) and X axis for error rate.

Expectations, plots, thoughts

Homopolymer Positions per Error Rate per Nucleotide :

- This will show where homopolymers are located in the sequence and how their positions change with different error rates.
- Given the increased error rate on homopolymers, I expect to see disruptions in CG positions of homopolymers as the error rate increases.
- A heatmap would show the positions of homopolymers for the 100 of iterations. The x-axis would represent sequence positions.

Explanations for the main function :

I generate a series of random DNA sequences based on given probabilities for each DNA base. The function uses the np.random.choice method to generate these. Each element is a randomly chosen base from BASES based on the probabilities P.

- Probability of G = 0.3
- Probability of C = 0.3
- Probability of A = 0.2
- Probability of T = 0.2

This function that introduces errors into a DNA sequence is the heart of the solution. It can be written using various logic, this is relatively simple implementation:

- It takes a DNA sequence and an error rate as input.
- Calculates the number of errors to introduce, based on the error rate and sequence length.
- Creates a probability distribution biased towards the end of the sequence to bias error positions. This is done by **linearly** increasing the probabilities from 0.5 to 1.5 across the sequence.
- Identifies homopolymer regions using helper function, and **doubles** their probability to increase likelihood of errors there.
- Randomly samples error positions from the position probability distribution, uses numpy's random.choice and samples without replacement so each position is picked only once (so no duplicate errors happening on the same positions).
- Randomly samples error types (distribution is equal in this case).
- For each error type, introduces errors at the chosen positions:
    - Insertions: Add a random base at the position
    - Mismatches: Change the base to a different random base
    - Deletions: Remove the base at the position
- Deletions are done last to avoid having to adjust positions after removals.
- The error positions and types are tracked in a dictionary to enable plotting the error position distribution later.

Thoughts on introduction of errors:

I've introduced a shift variable to account for the changes in sequence length due to insertions and deletions. After determining the positions and types of errors, I've sorted them based on positions to process them sequentially. As I introduce each error, I adjust the position using the shift and update the shift accordingly. This way the positions remain accurate even as the sequence length changes

Results - Dashboard vs Script

I am adding both the Dash dashboard and the script that builds plots.

These solutions are based on the same logics and dataset generation, but use slightly different plots, as Dash uses Plotly, which is not always the best for scientific plots.

Dashboard

DNA error rate dashboard allows you to explore the effects of introducing different error rates into randomly generated DNA sequences.

You can input custom error rates, number of sequences, and sequence length.

The app will then generate sequences with the specified error rates and visualise the impacts through requested plots.

The error distribution plot shows the probability distribution of error positions along the sequence length for insertion, deletion and mismatch errors.

The length distribution plot displays the sequence length distribution per error rate.

Higher error rates tend to introduce more insertions and deletions, increasing sequence length variation.

The GC content plot shows how increasing error rates slightly reduce average GC content. More errors break up long homopolymer runs of Gs and Cs.

The homopolymer ratios plot displays how higher error rates influence the proportional amount of homopolymers for each base type.

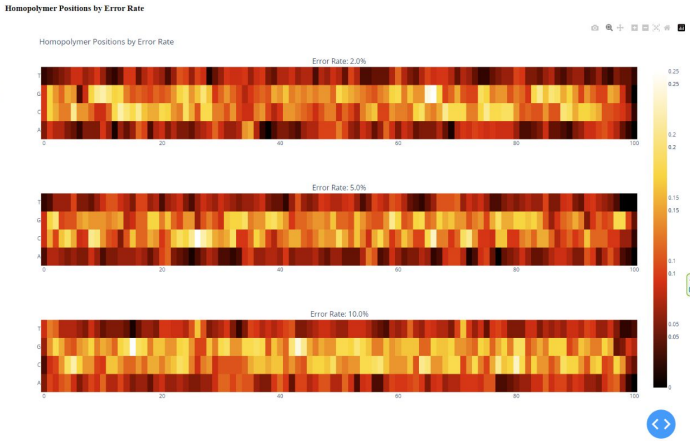
Finally, the heatmap visualises the frequency of homopolymer positions along the sequence for each error rate. Higher error rates decrease homopolymer frequency.

If you don't want to run the dahsboard using cmd “python dna_dash.py” command please open the Dash.mhtml file using Chrome (preview file)

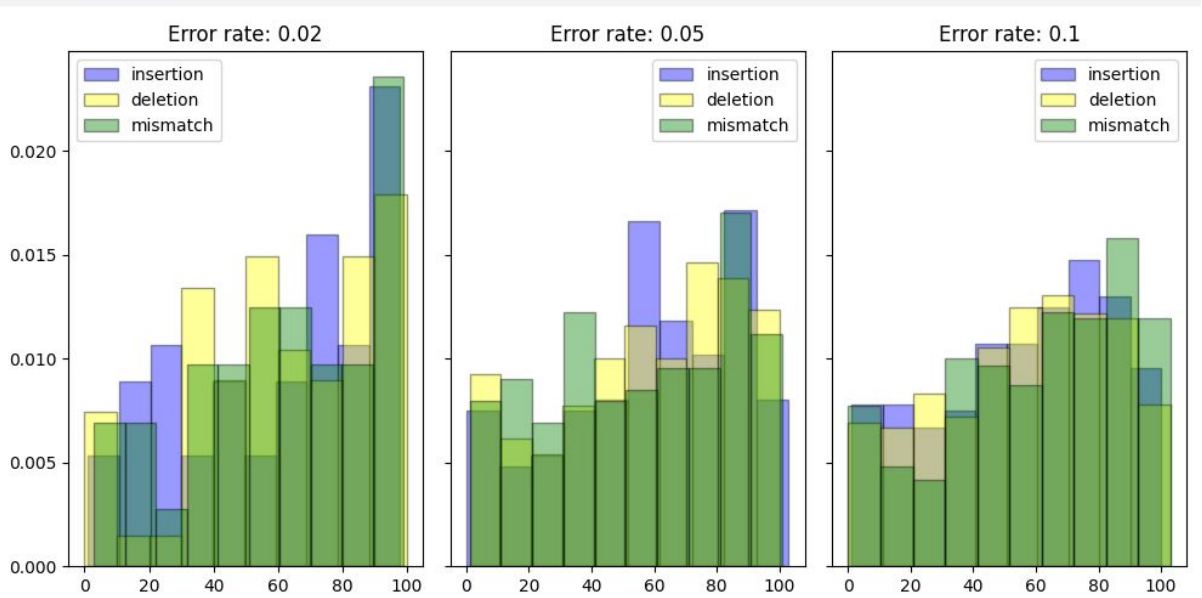
Dashboard

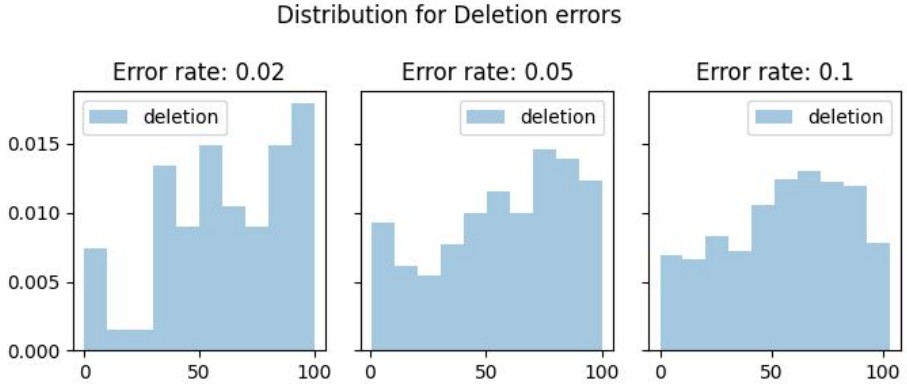

DNA Error Rate Dashboard
This dashboard allows interactive exploration of how error types and rates impact generated sequence properties like length, GC content and homopolymer ratios
Error Rates (comma-separated) 0.02 0.05 0.1
Number of Sequences 100
Amount of Bases 100
Generate Data and Update Plots
Insertion
Positions for Insertion errors
Error rate: 2.0%
Error rate: 5.0%
Error rate: 10.0%
Probability
Position in Sequence
Sequence length distribution per error rate
Number of Sequences
Sequence Length
GC Content vs. Error Rate
Average GC Content vs. Error Rate
Average GC Content (%)
Error Rate

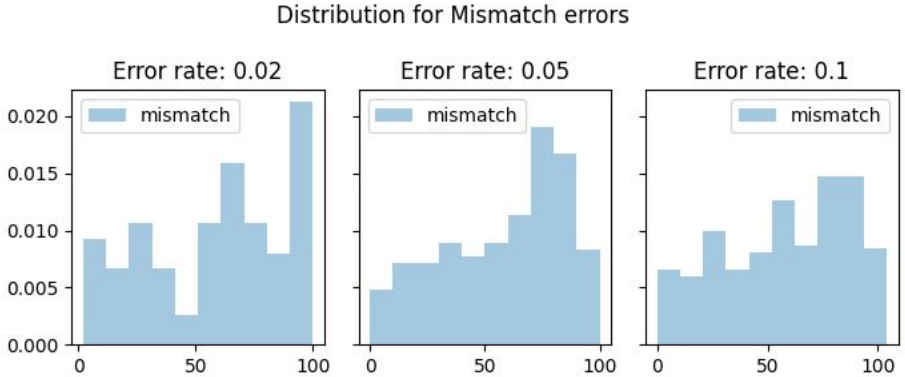

Homopolymer Ratios by Error Rate
Homopolymer Ratios by Error Rate
Homopolymer Ratio
Error Rate
Base
A
C
G
T


Homopolymer Positions by Error Rate
Homopolymer Positions by Error Rate
Error Rate: 2.0%
Error Rate: 5.0%
Error Rate: 10.0%

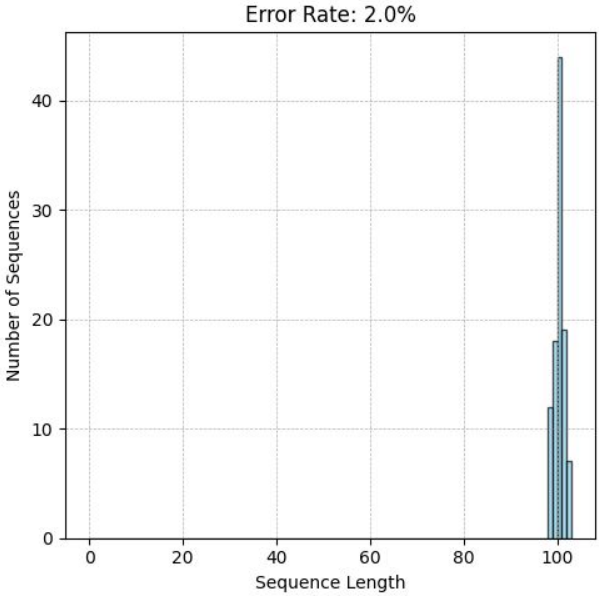
Script

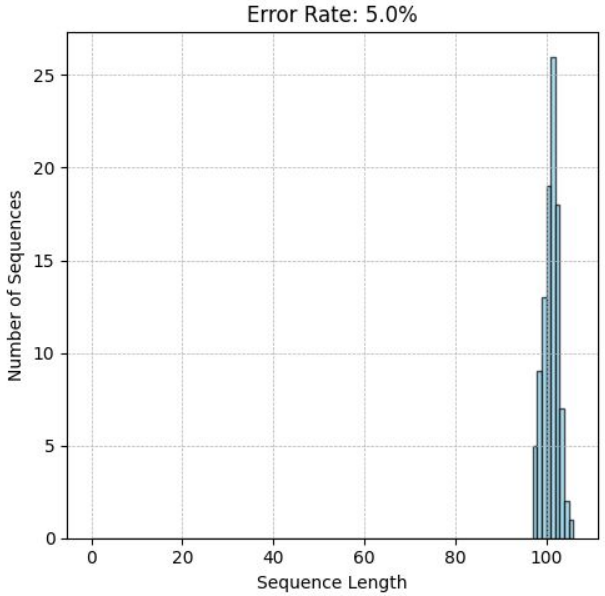

Error rate: 0.02
Error rate: 0.05
Error rate: 0.1
insertion
deletion
mismatch
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100

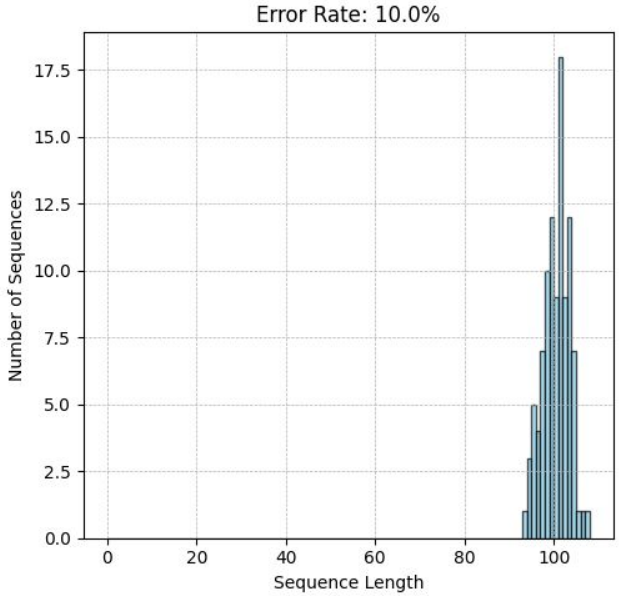




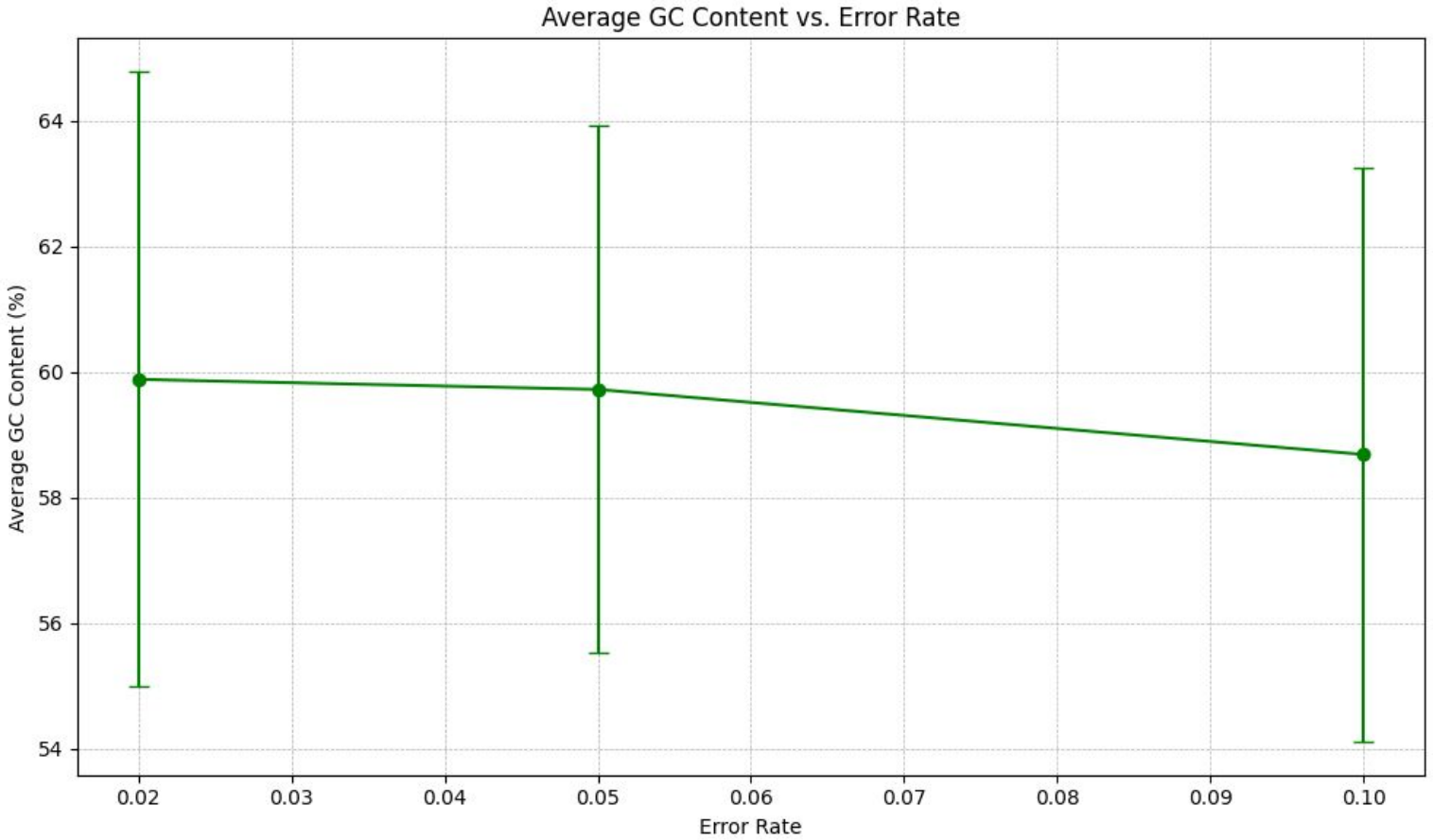
Error positions :
As more errors are introduced the gradient of errors increases towards the end of the sequence, with pronounced peaks at homopolymer positions.







Length of the sequences:
As more errors are introduced, the deletions and insertions are altering the length of the base sequence, so we see that the distribution becomes wider. As opposed to it, with lesser number of errors sequences length distribution is more homogeneous as lengths are closer to 100


Average GC Content vs. Error Rate
Average GC Content (%)
64
62
60
58
56
54
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
Error Rate

The initial sequence has a GC content of 60% (0.3 for C + 0.3 for G), so I expect more errors occurring at the respective sites, leading to slightly lower CG content

- **Insertions**:

When an insertion error occurs, a random nucleotide is added to the sequence. This could be any of the four bases (A, C, T, G). If the inserted base is either G or C, the GC content would increase, and if it's A or T, the GC content would decrease. Given the equal probabilities, there's an equal chance for all bases.

- **Deletions**:

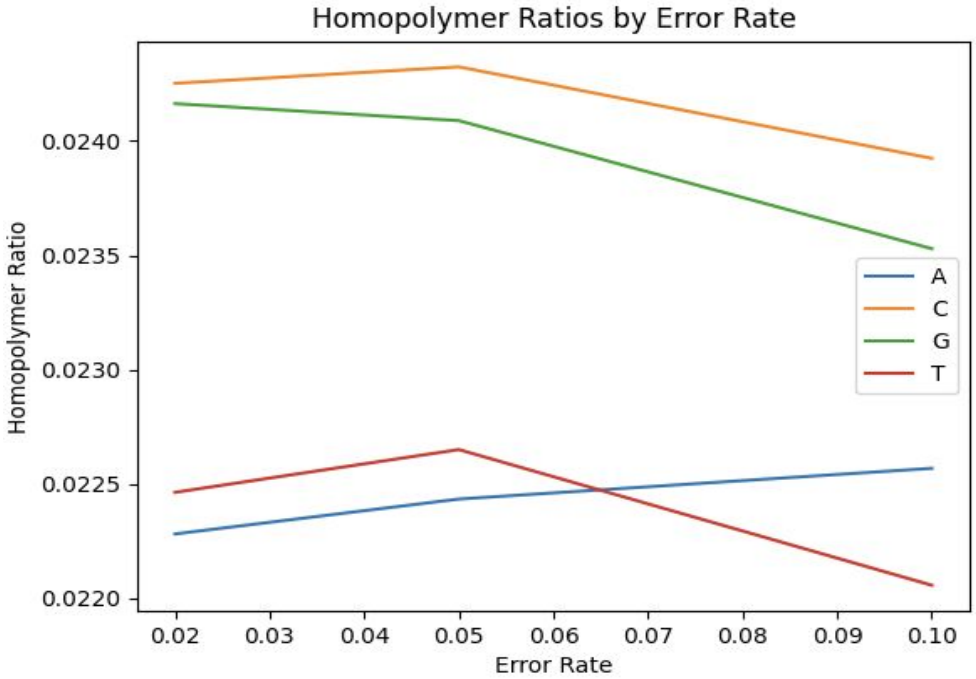
If a G or C is deleted, the GC content would decrease. If an A or T is deleted, the GC content would increase. The impact of deletions on GC content would depend on the original sequence (how many more homopolymers of C and G occurred, as this would increase errors on these sites further).

- **Mismatches**:

If a G or C is replaced with an A or T, the GC content would decrease. Conversely, if an A or T is replaced with a G or C, the GC content would increase.

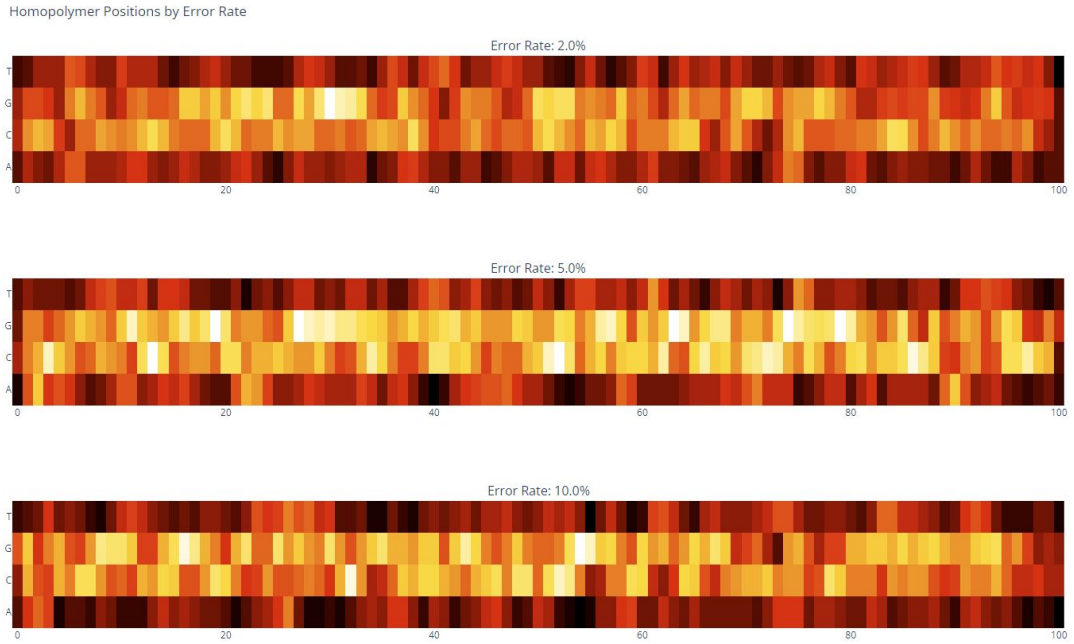
Overall Impact:

- Given the biases towards errors at the end of the sequence and within homopolymer regions, the GC content is fluctuating more in these areas.
- However, since the error rates provided (2%, 5%, and 10%) are relatively low compared to the sequence length, the overall GC content might not deviate drastically from the initial 60% depending on iteration. The exact deviation depends on the nature and distribution of the errors.



At bigger error rates, the ratios of C and G homopolymers decline. This is due to insertions, mismatches, and deletions disrupting these homopolymer stretches. A and T homopolymer ratios are a bit more variable, sometimes exhibit a slight uptick with increased error rates.

This variability is influenced by the nature of errors, with disruptions in C and G regions potentially increasing new A or T homopolymers. In essence, while C and G homopolymer ratios consistently decrease with escalating error rates, A and T ratios display a more nuanced and variable trend.



I consider the heatmap a good choice for this plot, as it shows what was discussed for the previous plot.

If we compare the 2% error rate vs the 10% error rate we would se more homogeneous colours at 2% breaking at10%, with more dark spots appearing.

Thank you for your time!